DOCUMENT RESUME

ED 132 162 TM 005 851

AUTHOR Kane, Michael T.; Moloney, James M.
TITLE Item Reliabilities for a Family of Answer-Until-Correct (AUC) Scoring Rules.
PUB DATE [Apr 76]
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 1976); Not available in hard copy due to marginal lebibility of original document

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS *Guessing (Tests); *Item Analysis; Mathematical Models; *Multiple Choice Tests; Probability; Raw Scores; Reliability; Response Mode; *Response Style (Tests); *Scoring Formulas; *Test Reliability; True Scores
IDENTIFIERS *Answer Until Correct Procedure

ABSTRACT

The Answer-Until-Correct (AUC) procedure has been proposed in order to increase the reliability of multiple-choice items. A model for examinees' behavior when they must respond to each item until they answer it correctly is presented. An expression for the reliability of AUC items, as a function of the characteristics of the item and the scoring rules for the item, is then derived from the model. The results indicate that the optimal scoring rule for an item depends strongly on the item's structure, and suggest general criteria for item selection when various scoring rules are used with the AUC procedure. (Author/MV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED132162

TM005 851 BEST COPY AVAILABLE

Item Reliabilities for a Family of Answer-Until-Correct (AUC) Scoring Rules

Michael Kane
SUNY, Stony Brook, N.Y.

James Moloney
SUNY, Brockport, N.Y.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, April 1976.

Multiple-choice items are typically scored on a two point scale; a correct answer gets a score of one, and an incorrect answer gets a score of zero. No attempt is made to discriminate among examinees who fail to answer the item correctly. Several variations on this procedure have been proposed in order to increase the efficiency of multiple-choice items (Coombs, Milholland & Wormer, 1956; Wang & Stanley, 1970; and Hanna, 1975). The aim of all of these procedures is to obtain more information from each item by allowing examinees who do not recognize the correct answer to indicate some level of partial knowledge.

The Answer-until-correct (AUC) procedure (Gillman and Ferry, 1972) employs a very simple and intuitive mechanism for making use of partial knowledge. When the AUC procedure is used on a test, the examinees are required to respond to each item until they choose the correct alternative for the item, and the score on the item is taken to be some function of the number of responses required to choose the correct alternative. The standard Zero-one (ZO) scoring procedure allows only one response to each item, and the additional responses required by the AUC procedure are intended to provide information on the level of partial knowledge for those examinees who cannot identify the correct alternative on their first try.

It is tempting to assume that the additional responses collected under the AUC procedure might have an effect similar to that of increasing the number of items, and we might reasonably

expect, therefore, that the AUC procedure would have a higher reliability than the ZO procedure. After responding incorrectly to an N-alternative item, the examinee responds to an item with the same stem but with the N-1 remaining alternatives.

The content of the item remains essentially the same, however, and the residual item is clearly not independent of the original item. The Spearman-Brown formula does not apply, and a more detailed analysis of examinee response behavior is necessary if we are to predict the effect of the AUC procedure on item reliability.

The few empirical studies that have been reported seem to indicate that the AUC procedure does produce some improvement in test reliability. Gillman & Ferry (1972) found that the split-halves reliability for AUC scoring was substantially higher than it was for ZO scoring. Hanna (1975) and Taylor, West, & Tinning (1975) report very slight increases in reliability when the AUC procedure was used. The empirical results, therefore do not provide strong support for the contention that the additional responses obtained under the AUC procedure are equivalent to additional items.

Some previous studies (Merwin, 1959; Brown, 1965; and Kane & Moloney, 1974) have indicated that the properties of the items determine whether the AUC procedure will have a higher, equal, or lower reliability than the ZO procedure. Since the

items used in empirical studies of the AUC procedure were not selected specifically for use with the AUC procedure, it is likely that some of the items used in these studies were more suitable for AUC scoring than for ZO scoring, and that some of the items were more suitable for ZO scoring than for AUC scoring.

This paper uses a simple model of examinee behavior to analyze the relationship between item characteristics and scoring rule (AUC or ZO) in determining item reliability. The results indicate that the two scoring rules require different item characteristics if the item reliability is to be maximized.

ASSUMPTIONS OF THE MODEL

The original version of the model was proposed by Horst (1930) to examine the effect of guessing on multiple-choice items. In analyzing the effect of guessing on items using the AUC procedure it is not necessary to make assumptions as strict as those used by Horst. In the discussion that follows it is possible to restrict ourselves to the following set of assumptions for a multiple-choice item with N alternatives:

(1) Each examinee recognizes n of the alternatives to be incorrect, and eliminates these alternatives from consideration, leaving a residual set of N-n alternatives.

(2) The correct alternative is never eliminated.

(3) The examinee selects alterratives from the residual set, randomly and without replacement, until the correct alternative is chosen.

The first assumption allows for the existence of degrees of partial information. An examinee who recognizes the correct answer eliminates all N-1 distractors, and selects the correct answer on his first try. An examinee who has no knowledge about the item selects randomly from the full set of N alternatives. An examinee who knows enough to recognize that n of the alternatives are incorrect eliminates those n alternatives, and selects randomly from the remaining N-n alternatives. The number of alternatives that can be eliminated defines the level of partial information for each examinee. This first assumption is the central assumption of Horst's model.

The second assumption places a restriction on the kind of misinformation allowed by the model. It states that no examinee makes the mistake of eliminating the correct alternative from consideration. However, examinees with a given level of partial information (defined here by the number of alternatives eliminated) may eliminate different subsets of the set of N-1 distractors. Examinees with a given level of partial information are not assumed to have the same information. If the alternatives are rank ordered according to their degree of plausibility, then examinees who can eliminate n of the alternatives do not necessarily eliminate the n worst alternatives, and the possibility of misinformation is allowed.

Horst's model does not allow for any misinformation. It makes the stronger assumption that the alternatives are uniquely

ordered from the correct alternative to the least plausible distractor, and that every examinee who eliminates n alternatives eliminates the n worst alternatives. As Lord & Novick (1968, P-312) have pointed out, this is the least plausible of Horst's assumptions. The model being used here does not require that the alternatives be uniquely ordered, but it does retain the assumption that the correct alternative is never eliminated.

The third assumption is a natural extension of Horst's model to the case where items are administered using the AUC procedure.

As we shall see in the next section, this simple model makes it possible to predict, for any monotone scoring rule, the proportion of item variance that is due to guessing. These results can then be used to select items for which the effect of guessing is minimized when a particular scoring rule is used. In subsequent sections we shall examine the implications of these results for two scoring rules that have been widely used.

IMPLICATIONS OF THE MODEL

When an examinee responds to a multiple-choice item using the AUC procedure, the following sequence of events is assumed to occur. The examinee eliminates n of the alternatives, and then responds by randomly selecting one of the remaining N-n alternatives. If this response is incorrect, the examinee makes

a random selection from the remaining N-n-1 alternatives. This process continues until the correct alternative is chosen.

Examinees who can eliminate all but k of the alternatives are said to be in the state C_k. Let α_k be the probability that a randomly selected examinee is in state C_k. The vector of probabilities,

$$\vec{\alpha} = (\alpha_1, \ldots, \alpha_N)$$

describes the structure of the item relative to some population, and will be called the item characteristic vector. Let S_j be the score given to any examinee who requires j responses to select the correct alternative. The vector,

$$\vec{S} = (S_1, \ldots, S_N)$$

specifies the numerical score to be assigned to any response sequence, and is called the scoring rule or score vector.

The probability that an examinee will respond correctly on his first try is just the probability of randomly selecting the correct alternative from a set of k alternatives. If X is a random variable designating the examinee's observed score, then the probability that an examinee in the state, c_k, answers correctly on the first try is given by:

$$P(X=S_1 | C_k) = \frac{1}{k}$$

The probability that an examinee in state C_k ($k \geq 2$) will require two responses on an item is equal to the probability of selecting a distractor on the first try, multiplied by the probability of selecting the correct alternative from the set of k-1 alternatives that remain after the first try:

$$P(X=S_2|C_k) = \frac{k-1}{k} \cdot \frac{1}{k-1} = \frac{1}{k} \qquad 2 \leq k$$

Since examinees in C_1 recognize the correct answer, the probability that they would require more than one response is zero. In general, the probability that an examinee in C_k will require j responses is given by

$$P(X=S_j|C_k) \begin{cases} =\frac{1}{k} & \text{for } j \leq k \\ =0 & \text{for } j > k \end{cases} \qquad (1)$$

The observed scores for the examinees in any state have a uniform distribution over the set of possible scores for that state.

The probability, Π_j, that an examinee, chosen randomly from the population, will require j responses to select the correct alternative is:

$$\Pi_j = \sum_{k=1}^{N} P(X=S_j|C_k) \cdot P(C_k) = \sum_{k=j}^{N} \frac{1}{k} \alpha_k \qquad (2)$$

The observed proportion of examinees requiring j responses to answer an item may be substituted for Π_j in Equation (2) in order to estimate $\tilde{\alpha}$.

The expected value of the observed score over the population of examinees is:

$$E(X) = \sum_{j=1}^{N} S_j \Pi_j = \sum_{j=1}^{N} S_j \sum_{k=j}^{N} \frac{1}{k} \alpha_k$$

Changing the order of summation, we have:

$$E(X) = \sum_{k=1}^{N} \frac{\alpha_k}{k} A_k \quad \text{where } A_k = \sum_{j=1}^{k} S_j \tag{3}$$

Similarly, the expected value of X^2 is:

$$E(X^2) = \sum_{k=1}^{N} \frac{\alpha_k}{k} B_k \quad \text{where } B_k = \sum_{j=1}^{k} S_j^2 \tag{4}$$

From Equations (3) and (4), the observed score variance for an item is given by:

$$\sigma^2(X) = \sum_{k=1}^{N} \frac{\alpha_k}{k} B_k - \left[\sum_{k=1}^{N} \frac{\alpha_k}{k} A_k \right]^2 \tag{5}$$

The true score, T_k, for all examinees in state, C_k, is just the expected value of X, given C_k:

$$T_k = \sum_{j=1}^{N} S_j \, P(X=S_j | C_k) = \frac{1}{k} A_k. \tag{6}$$

The true score variance, over the population of examinees, is

$$\sigma^2(T) = \sum_{k=1}^{N} \left(\frac{A_k}{k}\right)^2 \alpha_k - \left[\sum_{k=1}^{N} \left(\frac{A_k}{k}\right) \alpha_k \right]^2 \tag{7}$$

The error for any examinee is the difference between his observed score on the item and his true score. The variance in E over independent trials, for an examinee in state, C_k, is the

observed score variance for examinees in C_k:

$$\sigma_k^2(E) = \sum_{j=1}^{k} S_j^2\left(\frac{1}{k}\right) - \left[\sum_{j=1}^{k} S_j\left(\frac{1}{k}\right)\right]^2 = \frac{B_k}{k} - \left(\frac{A_k}{k}\right)^2$$

Therefore, the error variance for an examinee is a function of the state of partial knowledge of the examinee, C_k. Those examinees who know least about the item do the most guessing. Examinees who recognize the correct answer never guess, and the error variance for C_1 is zero:

$$\sigma_1^2(E) = 0$$

The error variance over the population is just a weighted average of the error variances for the particular states:

$$\sigma^2(E) = \sum_{k=1}^{N} \alpha_k \sigma_k^2(E) = \sum_{k=1}^{N} \frac{B_k}{k} \alpha_k - \sum_{k=1}^{N} \left(\frac{A_k}{k}\right)^2 \alpha_k \qquad (8)$$

From Equations (5), (7), and (8) it follows that:

$$\sigma^2(X) = \sigma^2(T) + \sigma^2(E)$$

The observed score variance is the sum of the true score variance and the error variance. Furthermore, it can easily be shown that the errors are uncorrelated with true scores and have an expected value of zero. With an appropriate definition of true score, then the model is a special case of the classical test theory model.

The item reliability can now be definfed as the ratio of the true score variance to the observed score variance:

$$R = \frac{\sigma^2(T)}{\sigma^2(X)} = \frac{\sum_{k=1}^{N} \left(\frac{\Lambda_k}{k}\right)^2 \alpha_k - \left[\sum_{k=1}^{N} \left(\frac{\Lambda_k}{k}\right)\alpha_k\right]^2}{\sum_{k=1}^{N} \left(\frac{B_k}{k}\right)\alpha_k - \left[\sum_{k=1}^{N} \left(\frac{\Lambda_k}{k}\right)\alpha_k\right]^2} \quad , \alpha_1 < 1.0 \qquad (9)$$

The assumption that $\alpha_1 < 1.0$ is necessary to assure that the observed score variance is different from zero. Because guessing is the only source of error being considered, Equation (9) is an upper bound for the usual indices of stability, equivalence, and internal homogeneity.

The item reliabilities given by Equation (9) depend on both the vector of probabilities, $\vec{\alpha}$, that describes the structure of the item relative to some population and the score vector, $\vec{S}$, that specifies the item score as a function of the number of responses needed to select the correct alternative.

SCORE VECTORS

Where the AUC procedure has been used (Gillman & Ferry, 1972; Hanna, 1975; and Taylor et. al., 1975), the scoring rule has had the following simple form:

$$S_k = N - k \qquad 1 \leq k \leq N$$

where k is the number of responses needed to choose the correct alternative. Because Equation (9) is invariant under linear transformations of the scoring rule, we will use an equivalent but more convenient form of this rule:

$$S_k = k \qquad 1 \leq k \leq N \qquad (10)$$

Although it is possible to examine the reliability of items for any monotone scoring rule for the AUC procedure by appropriate substitutions in Equation (9), the rule specified by Equation (10) will be taken as defining the AUC scoring rule for the remainder of this paper.

For the ZO procedure, the examinee responds only once to each item, and an appropriate scoring rule is given by:

$$S_1 = 1$$

$$S_k = 0 \quad \text{for } 2 \leq k \leq N \tag{11}$$

When the ZO procedure is used, we do not know how many responses an examinee would have required to choose the correct alternative, but this is not a serious problem here because the same score is assigned to all examinees who require more than one response to choose the correct alternative.

TWO SPECIAL CASES

In this section, two special cases for the form of the item characteristic vector, $\vec{\alpha}$, will be discussed because they illustrate the content of Equation (9), and because they will be useful in our subsequent discussion.

First, we consider N-alternative items for which each examinee can eliminate at least N-2 of the alternatives. This implies that there is only one effective distractor for each

examinee but it does not imply that the same distractor is effective for all examinees. The item characteristic vector has the form:

$$\tilde{\alpha} = (\alpha_1, \alpha_2, 0, \cdots, 0)$$

Substituting these values into Equation (9) and simplifying the resulting expressions, we get the same result for both the AUC scoring rule and the ZO scoring rule:

$$R = \frac{\alpha_1}{\alpha_1 + 1} \tag{12}$$

It is not surprising that both scoring rules lead to the same expression for R because, in this case, the item is effectively a two alternative item for each examinee, and the AUC scoring rule is then a linear transformation of the ZO scoring rule.

The second case to be discussed assumes that all of the examinees fall into one of two states. They can either eliminate all of the distractors and choose the correct alternative on their first response, or they cannot eliminate any of the distractors and guess randomly until they choose the correct alternative. The item characteristic vector corresponding to these assumptions is:

$$\tilde{\alpha} = (\alpha_1, 0, \cdots, 0, \alpha_M) \tag{13}$$

Item reliabilities for AUC-scored items that satisfy these assumptions can be derived from Equations (9), (10), and (13).

and the result after algebraic simplification is given by:

$$R = \frac{(N-1)\alpha_1}{(N-1)\alpha_1 + \frac{1}{3}(N+1)} \qquad (14)$$

Reliabilities for ZO-scored items can be derived similarly from Equations (9), (11), and (13):

$$R = \frac{(N-1)\alpha_1}{(N-1)\alpha_1 + 1} \qquad (15)$$

For this case, then, the reliability for ZO scoring is greater than it is for AUC scoring. As Horst's original paper (1930) showed, the item characteristic vector given in Equation (13) minimizes the effect of guessing for multiple choice items scored in the traditional way. The probability that an examinee who does not know the correct answer will choose the correct alternative on the first response is a minimum, 1/N, when the examinee can not eliminate any of the distractors.

The reason for the superiority of ZO scoring when all of the distractors are equally plausible is obvious. Under this condition only the first response supplies any information about the examinees. All of the examinees who have any information on the item, answer it correctly on their first response (along with some examinees who guess correctly on their first response). All responses on an item, after the first, are guesses and supply no reliable information on the examinee. In this case, the additional responses required by the AUC procedure do

nothing but generate noise.

ITEMS WITH THREE ALTERNATIVES

The special cases analyzed in the previous section examined some of the implications of Equation (9), but the full content of that result is still somewhat obscure. In this section, the relationship between item reliability and item characteristics will be examined for items with three alternatives. The results in this section also apply to items with more than three alternatives for which each examinee is able to eliminate at least N-3 of the alternatives. Such items represent the simplest cases where there is a difference between the ZO scoring rule and the AUC scoring rule.

Insert Table 1 about here

Table 1 presents the values of R for AUC scoring and ZO scoring for items with three alternatives (increments of 0.1 are used for α_1 and α_2, and $\alpha_3 = 1.0 - \alpha_1 - \alpha_2$). The cells in Table 1, for given values of α_1 and α_2, consist of two entries. The first entry is the AUC reliability as computed from Equations (9) and (10), and the second entry is the ZO reliability as computed from Equations (9) and (11). The broken line in Table 1 separates it into two parts. In the upper part, the reliability for ZO scoring is larger than it is for AUC scoring. In the bottom part, the reliability for ZO scoring is less than or equal to

the reliability for AUC scoring

The top row in Table 1 is an example of the second of the special cases discussed in the last section. If α_2 is zero, then all of the examinees fall into one of two categories; they either know the correct answer or they have no information about the item and guess randomly. ZO reliability is uniformly higher than the AUC reliability under these conditions.

For the last entry in each column, α_3 is zero, and the item is effectively a two-alternative item. As shown previously, the reliabilities for AUC scoring and ZO scoring are identical in this case.

For a fixed value of α_1, a column in Table 1, the ZO reliability is a maximum when α_2 equals zero, and decreases approximately linearly for increasing values of α_2, while the AUC reliability generally increases at first, reaches a maximum, and then decreases as a function of α_2. The structure of Equation (9) and the details of Table 1 reflect the complex relationship between item characteristics and the scoring rule. For any number of alternatives and any item characteristic vector, the two scoring rules can be compared by evaluating Equation (9). When items with more than three alternatives are examined in this way, the relationship between AUC reliability and ZO reliability becomes quite complicated. There are,

however, two general characteristics of the AUC reliability that are very clear in Table 1, and that apply to items with any number of alternatives.

First, the possible improvements to be derived from the substitution of AUC scoring for ZO scoring are most pronounced when α_1 is relatively small. When α_1 is large, a large proportion of the examinees know the correct answer, and any procedure for assessing partial information will affect relatively few examinees. It is therefore not surprising that the substitution of AUC scoring for ZO scoring has little effect on item reliability when α_1 is large.

Second, for a given value of α_1, the AUC reliability tends to be highest when both α_2 and α_3 are appreciably different from zero. For items with N alternatives, the AUC reliability tends to be highest when $\alpha_2, \alpha_3, \cdots \alpha_N$ are all appreciably different from zero. It is only when the examinees differ in their levels of partial information that they can be discriminated on the basis of partial information. To the extent that examinees do not differ in their levels of partial knowledge, the additional responses called for by the AUC procedure do nothing but increase the variance due to guessing.

The AUC procedure has generally been associated with the linear scoring rule given in Equation (10), and we have called this rule the AUC scoring rule. It is evident, however, that the class of admissible scoring rules for the AUC procedure includes all monotone non-increasing (or monotone non-decreasing) functions of k, where k is the number of responses needed to answer the item.

In this section, some of the possible scoring rules for the AUC procedure are examined briefly. To make this task manageable, it is convenient to focus on the family of scoring rules having the following general form:

$$S_k = k^c \tag{16}$$

where c is a constant, other than zero. Even when attention is restricted to this simple type of scoring rule, there is an infinite number of candidates. Using Equations (9) and (16), it is possible to obtain a general expression for the item reliability as a function of the item characteristic vector and the constant, c. The resulting expression is, however, too complicated to be directly informative, and we shall proceed by examining special cases.

The two scoring rules discussed in the preceding sections of this paper are special cases of Equation (16). The AUC scoring rule is obtained by letting c be equal to one. The ZO scoring rule is obtained by taking the limit of Equation (16) as c goes to minus infinity.

Insert Figure 1 about here

The relationship between item reliability, item characteristics, and scoring rule is illustrated in Figure 1 for four values of c. Figure 1 presents item reliabilities for 5-alternative items with $\alpha_1 = 0.1$ and $\alpha_2 = \alpha_3 = \alpha_4$. As mentioned earlier, the differences between the scoring rules tend to be greatest when α_1 is relatively small. The second restriction, setting $\alpha_2 = \alpha_3 = \alpha_4$, is introduced to simplify the presentation. With these restrictions, the item characteristic vector has one degree of freedom, represented by α_5. The curves in Figure 1 represent the item reliabilities for four scoring rules as functions of α_5. Each scoring rule is identified by the appropriate value of c. Thus, the inverse scoring rule, $S_k = 1/k$, is identified by c = -1.0, and the square root scoring rule is identified by c = 0.5.

The most obvious and most important point illustrated by Figure 1 is the fact that none of the scoring rules optimizes the reliability for all items. When $\alpha_5 \leq 0.3$, the AUC scoring rule has the highest reliability of the four rules under consideration here. When α_5 is close to 0.9, the ZO scoring rule has the highest reliability. The inverse scoring rule and the square root scoring rule have the highest reliability for the intermediate values of α_5. In general the best scoring rule for an item depends on the characteristics of the items.

CONCLUSIONS

The total number of responses obtained for any set of items will be greater for the AUC procedure than it is for the ZO procedure. Gillman & Ferry (1972) have suggested that these additional responses will tend to increase the reliability of a test by effectively lengthening it. The results reported here can be interpreted as an analysis of the effect of these additional responses on item reliability, where ZO scoring provides a standard of comparison.

The accuracy of the predictions that have been made here will depend on how closely the assumptions of the model relect actual test-taking behavior. Horst's model is clearly more realistic than the traditional model for guessing, and the version used here has relaxed the least plausible of Horst's assumptions -- that the alternatives can be uniquely ordered. It is expected then that the present model provides a relatively realistic description of examinee behavior.

The model predicts that the effect of guessing on item reliability depends on both the item characteristics and the scoring rule. The empirical research that has compared AUC scoring to ZO scoring has ignored the effect of item characteristics. This failure to control item characteristics may account for the fact that AUC scoring has not consistently produced improvements in reliability when it has been compared to ZO scoring.

The admission of additional responses for the AUC procedure increases the observed score variance. If the AUC procedure produces a larger proportional increase in true score variance than in error variance, the use of this procedure will increase reliability. Otherwise, it will tend to decrease item reliability. This paper has produced some model-based predictions about the item characteristics that determine which of these conditions will hold.

Equation (9) makes it possible to select items that will minimize the effect of guessing when the AUC procedure is used. Alternately, items can be analyzed to judge whether any improvement in reliability can be expected if AUC scoring is substituted for ZO scoring.

REFERENCES

Brown, J. *Multiple Response Evaluation of Discrimination.* British Journal of Mathematical and Statistical Psychology, 1965, 18, 125-137.

Coombs, C.H., Milholland, J.E., & Wormer, F.B. *The Assessment of Partial Knowledge.* Educational and Psychological Measurement, 1956, 16, 13-37.

Gilman, D. & Ferry, P. *Increasing Test Reliability Through Self-Scoring Procedures.* Journal of Educational Measurement. 1972, 9, 205-207.

Hanna, G. *Incremental Reliability and Validity of Multiple-Choice Tests with an Answer-Until-Correct Procedure.* Journal of Educational Measurement, 1975, 12, 175-178.

Horst, P. *The Chance Element in the Multiple-Choice Test Item.* Journal of General Psychology, 1932, 6, 209-211.

Kane, M. & Moloney, J. *The Effect of SSM Grading on Reliability when Residual Items have no Discriminating Power.* Paper presented at the meeting of the American Educational Research Association, Chicago, 1974.

Lord, F. & Novick, M. *Statistical Theories of Mental Test Scores.* Reading, Mass., Addison-Wesley, 1968.

Merwin, J.C. *Rational and Mathematical Relationship of Six Scoring Procedures Applicable to Three-Choice Items.* Journal of Educational Psychology, 1959, 50, 153-161.

Taylor, J., West, D., & Tinning, F. *An Examination of Decision-Making Based on a Partial Credit Scoring System.* Paper presented at the meeting of the National Council on Measurement in Education, Washington, D.C., 1975.

Table 1
AUC and ZO Reliabilities for Items with 3 Alternatives

α_2	α_1 0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0.00									
AUC	.130	.231	.310	.375	.429	.474	.512	.545	.574
ZO	.167	.286	.375	.444	.500	.546	.583	.615	.643
0.10									
AUC	.155	.248	.323	.383	.433	.474	.506	.528	.474
ZO	.166	.277	.360	.424	.474	.511	.536	.542	.474
0.20									
AUC	.176	.262	.331	.387	.432	.466	.485	.444	
ZO	.163	.267	.344	.402	.444	.472	.480	.444	
0.30									
AUC	.192	.272	.335	.385	.423	.443	.412		
ZO	.158	.255	.326	.377	.412	.427	.412		
0.40									
AUC	.202	.276	.333	.375	.398	.375			
ZO	.152	.241	.306	.350	.375	.375			
0.50									
AUC	.207	.274	.322	.348	.333				
ZO	.144	.226	.283	.320	.333				
0.60									
AUC	.204	.261	.294	.286					
ZO	.133	.208	.258	.286					
0.70									
AUC	.190	.231	.231						
ZO	.121	.189	.231						
0.80									
AUC	.158	.167							
ZO	.107	.167							
0.90									
AUC	.091								
ZO	.091								

FIGURE 1

ITEM RELIABILITIES FOR FIVE-ALTERNATIVE ITEMS (WITH $\alpha_2 = \alpha_3 = \alpha_4$, AND $\alpha_1 = 0.1$) WITH SCORING RULES $S_K = K^C$